Project Proposal Group members: Zhaohong Jin, Shu Xu, Yiming Xu

1. What is the goal of the project/idea?

The goal of this project is to study how to protect user privacy in social networks such as Facebook that use image recognition networks to match identities. When users upload photos to Facebook, they are automatically tagged with their names and are exposed to risks such as identity thefts or spam advertisements. To solve this, we will examine adversarial attacks, a trending machine learning research nowadays, on image recognition algorithms to fool them into identifying a user face as something else. Adversarial attacks work by treating the image recognition networks as a black box and applying optimization techniques to generate fooling images. The end product is an identity protection analysis that automatically finds the most efficient way to add human unnoticeable modification to the user pictures such that they are robust against the strongest neural nets.

2. What data will you use/need to collect?

In this project, we are going to use PubFig as our dataset. PubFig is a real-world dataset consisting of pictures of 200 celebrities, with an average of 300 pictures per person. We chose PubFig for the following reasons. First, the dataset labels the name of each person, which is the personal identity information we are trying to protect. Second, these face images have different lighting, background, and facial expressions, so it's good simulation of what people would post on social networks.

3. What current/future tools from class do you plan to use?

We will use Tensorflow framework to code our model and to visualize our image. For our face recognition model, we planned to use a pre-trained Google VGG-Net. This corresponds to the neural net chapter from the class. In addition, to create adversarial images, we will examine various optimization techniques such as saliency map, Fast Gradient Sign Method, Black Box attack, White Box attack from different sources(papers, tutorials, etc) and developing our own ways of attack. This requires optimization knowledge such as defining a proper loss function, finding solutions using gradient descent, and making visualizations.

4. Provide a week-by-week plan of attack: one sentence for each of the next 6 weeks that describes what you plan to accomplish.

Week 1&2 Finalize proposal, prepare dataset (SX), literature review and EDA (All).
Week 3 (April 8) Build the facial classification neural network(SX)
Week 4 (April 15) Generate adversarial image using current methods(ZJ), evaluate the privacy impact(YX), visualize results(SX) and understand how it works (All).
Week 4 (April 15) Explore other methods/propose our own, evaluate and visualize(All).
Week 6 (April 29) Gather results, draw conclusions and prepare for presentation (All)

5. Assign each of the group members to one or more of the task from Step 4.

Zhaohong Jin: attacking methods selection, generate adversarial image, code infrastructure.
Shu Xu: prepare dataset, build base model, explore privacy impact, visualize results
Yiming Xu: perform exploratory data analysis, evaluate results, explore privacy impact
All: literature review, understand the mechanism, try and propose new methods and visualize them, gather results, draw conclusions and prepare for presentation.